FRONTIERS IN PSYCHOLOGY

REVIEW ARTICLE
published: 15 October 2010
doi: 10.3389/fpsyg.2010.00167

A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating

Michalis P. Michaelides*

Department of Education Sciences, European University Cyprus, Nicosia, Cyprus

Edited by:
Lawrence Rudner, Graduate Management Admission Council, USA

Reviewed by:
Weihua Fan, University of Houston, USA
Elizabeth Stone, Educational Testing Service, USA

****Correspondence:***
Michalis P. Michaelides, Department of Education Sciences, European University Cyprus, 6, Diogenes Street, Engomi, P.O. Box 22006, 1516 Nicosia, Cyprus.
e-mail: m.michaelides@euc.ac.cy

Many studies have investigated the topic of change or drift in item parameter estimates in the context of item response theory (IRT). Content effects, such as instructional variation and curricular emphasis, as well as context effects, such as the wording, position, or exposure of an item have been found to impact item parameter estimates. The issue becomes more critical when items with estimates exhibiting differential behavior across test administrations are used as common for deriving equating transformations. This paper reviews the types of effects on IRT item parameter estimates and focuses on the impact of misbehaving or aberrant common items on equating transformations. Implications relating to test validity and the judgmental nature of the decision to keep or discard aberrant common items are discussed, with recommendations for future research into more informed and formal ways of dealing with misbehaving common items.

Keywords: test equating, common items, item parameter estimates

INTRODUCTION

Large-scale testing programs provide scores for individual achievement or ability, and aggregate scores for examinee groups – in the case of educational tests, for schools, districts, or states. Scores are often derived from alternate versions of a test administered over different occasions. While this is a way to guard against the overexposure of the content and ensure the security of the test, it creates the problem of score interchangeability. Alternate test forms will be differentially difficult for examinees; however for the sake of fairness, it should not matter to them which test form they take (Lord, 1980). Test equating methods are statistical adjustments that establish comparability between alternate forms built to the same content and statistical specifications by placing scores on a common scale (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999; Kolen and Brennan, 2004).

TEST EQUATING USING COMMON ITEMS

Various designs are available to conduct equating between test forms, such as the random groups, or the single group designs. In this paper, the focus is in the common-item non-equivalent groups design, or "Non-Equivalent groups with Anchor Test" (NEAT) design (von Davier et al., 2004), where two forms are equated through a subset of common items embedded in both forms. These anchor items can be internal and contribute to the examinee total test score; or they may be external to the test, not contributing to the total score and typically administered as a separate section of the test (Kolen and Brennan, 2004). Performance on the common items is used to establish the linking relationship between the groups taking the alternate forms.

A key assumption made when an equating is performed under the NEAT design is that the statistical properties of the common items that operate as anchors are stable across forms; when two groups respond to two alternate forms, the common items must function similarly in both forms (Hanson and Feinstein, 1997; Wainer, 1999). If two groups of examinees respond differently to the same item, then that item might not be appropriate to be included in the equating process. If an equating item demonstrates a large change in its difficulty index or its item response theory (IRT) parameter estimates, it raises suspicion, and calls for inspection. Experts seek to determine possible reasons for why the item functions differentially. They can speculate whether the differential performance is related to the purpose of measurement, i.e., if it reflects a true change in the proficiency of the examinee cohorts, or if it is due to irrelevant factors, such as a change in the position of the item in the test form. It may be kept in the anchor set, or it may be discarded and treated as a regular, non-common item, as if there was no connection between the item in the first and the item in the second form. Inclusion or exclusion of an item from the equating pool is a matter of judgment and impacts the equating function. It should be noted that any effects may become more pronounced in chain equating designs where a series of test forms are equated sequentially.

IRT ASSUMPTIONS AND PARAMETER INVARIANCE

Test equating is a component of a larger, cyclical, measurement process that involves test development, administration, analysis, scoring, reporting, and evaluation (Hattie et al., 1999). When a model, such as any IRT model, provides the conceptual measurement framework for this process, the results depend on how well the data fit that model.

Parameter invariance, the property of IRT item parameter estimates to remain unchanged across various groups of examinees, and ability estimates to remain invariant across groups of items, gives IRT its applicability and usefulness (Lord, 1980; Allen et al., 1987; Linn, 1990; Hambleton et al., 1991). According to parameter invariance, if the IRT model fits the data perfectly, then parameters will be invariant across administrations, except for sampling

fluctuations that introduce random error in the responses of examinees. In that case, the changes in the behavior of item parameter estimates would follow a systematic pattern depending on the changes in the size and proficiency of the different examinee groups. There is a further issue in the IRT framework: invariance does not imply mathematical identity, since an IRT model is based on an arbitrary latent scale. To resolve this scale indeterminacy when calibrating response data, the latent ability θ is typically required to be normally distributed with mean 0 and standard deviation 1; subsequently, parameter estimates across administrations will be invariant only up to a set of linear transformations (Lord, 1980; Rupp and Zumbo, 2006).

Item response theory makes strong assumptions and its promise for invariance depends on the degree that the model assumptions, and particularly unidimensionality, hold (Miller and Linn, 1988). Violation of the unidimensionality assumption is potentially a major source of problems for IRT equating (Skaggs and Lissitz, 1986). Suppose estimation of item parameters using data from two different groups of examinees yields different item parameters, or equivalently, that two test-characteristic curves exist for the same test.[1] Examinees in the two groups who obtain equal ability estimates from the model would have different probabilities of answering items right. According to Lord (1980), if a test discriminates between examinees of the same ability, it actually measures a dimension other than the intended ability. Unidimensionality is not defensible and the assumption of invariance is then dubious. Even though unidimensional models do not fit to assessment settings where multiple proficiencies are engaged simultaneously in tasks of interest, common practice employs unidimensional models to analyze such tests.

To make an argument for unidimensionality, a dominant component affecting test performance would suffice (Hambleton et al., 1991). However, empirical findings have been consistent in pointing to departures from unidimensionality that are usually large, thus casting doubt on the underlying assumptions and the inferences drawn from the model.

Large changes or differences in instructional experience may be needed to produce practically significant violations of assumptions, and the effects may be very specific and limited to few items of a test (Linn, 1990). However, IRT models are approximations at best. And thus anomalous item behavior may not be ruled out, even if models perform sufficiently well to justify their continued use. The problem becomes more profound in test equating when equating functions are derived from item parameter estimates of the common items. It is not unusual to come across a few common items that do not follow the behavior of the majority of the common items. Those misbehaving items are checked for possible reasons that caused their anomalous behavior. It is then a judgmental decision whether to keep them or discard them from the equating pool. An obvious explanation might exist for that behavior, such as an inadvertent change in the precise wording or formatting of an item, or differential instructional emphasis on the content of an item, or there may be no immediate compelling reason triggering that behavior.

In this paper, a review of the existing literature reveals that characteristic indices of items change and item parameters are not invariant across different administrations, even though IRT models that are usually fitted to test data rest on the assumption of parameter invariance. Since item parameter estimates for common items are utilized to generate equating transformations, the accuracy of the equating will be impacted by such drift.

CURRICULAR EFFECTS ON ITEM PARAMETER ESTIMATES

Many empirical studies address the adequacy of IRT models by examining whether parameter invariance or unidimensionality hold using real or simulated data. Miller and Linn (1988) examined the effect of differential instructional coverage on item characteristic functions. They grouped students who participated in the Second International Mathematics Study into curriculum clusters based on their teacher's ratings of their opportunity to learn the content of each of the test items during the previous year. Item characteristic curves for the arithmetic and algebra items for each of the curriculum clusters were compared. Large differences were detected between the curves, indicating that item parameters were influenced by variations in opportunity to learn.

Masters (1988) provided evidence for differential item performance caused by the opportunity to learn particular content in high- versus low-level mathematics classes. For example, items on content that one group had more opportunity to learn had different difficulty parameters when separate calibrations were made for each group. If the two groups' responses were calibrated simultaneously, the difficulty parameter would fall between the previous two values, and the discrimination parameter would be higher if the group that had more opportunity to learn was on average of higher ability.

Content analysis may help explain findings of item parameter drift (Linn, 1990). Bock et al. (1988) found differential linear drift of the item location parameters in items of a College-Board Physics Achievement Test over 10 years. They associated the direction of the drift with the content of the items in a pattern that reflected a changing emphasis in secondary school physics curricula. Considering a lack of substantial drift in English items over that same time period, they attributed the noticeable drift in physics items to the greater likelihood of change in physics curricula. Among 29 mechanics items, 11 that referred to basic concepts became easier over time, while the difficulty of 10 other items less related to basic concepts increased. Their evidence suggests a decreased emphasis on advanced and specific topics, which may reflect a back-to-basics approach in physics textbooks. A pair of mechanics items on the difference between mass and weight, one of which used metric and the other English units, exhibited drift in opposite directions. The cases moved in a direction that reflected the introduction of metric units at the end of the 1970s, i.e., items with metric units became easier as the new units became more familiar, while the opposite happened with the items referring to English units that were gradually phased out of the curriculum. Apart from systematic item location drift, Bock et al. (1988) observed occasional anomalies in some items. They suggested that such cohort-specific effects are

[1]Because item parameters are invariant only up to a linear transformation of the ability scale, item parameter estimates obtained using different examinee groups would need to be transformed to be placed on a common scale. This illustration refers to differences in item parameter estimates that remain after such rescaling has been carried out.

unexpected in large nationwide samples but may reflect special attention given to some topics by the media or publications accessible to physics teachers.

Sykes and Fitzpatrick (1992) classified a large number of items from consecutive administrations of a professional licensure examination into four content categories. In one of the four categories, they detected a significantly greater drift of Rasch *b* parameter estimates. They hypothesized that the "differential change in *b* values is attributable to shifts in curriculum emphasis, with the most pronounced shift occurring for the content covered in this category" (p. 210).

Much research on item parameters emerged from studies of customized tests and the validity of estimates drawn from actual or simulated customizations. In the early 1980s, national tests were often customized by local authorities and adjusted to extract national normative scores for the local examinees. The validity of such inference has been questioned. Consistent findings indicate that item calibrations are not invariant across samples. Yen et al. (1987) present a case where the IRT difficulty parameter estimates in the national calibration of a mathematics test changed systematically at the local level; in a local calibration the measurement items were relatively more difficult, while the numeration items were relatively easier, suggesting that different local curricular and/or instructional characteristics influenced parameter estimates.

Allen et al. (1987), Linn (1990), Way et al. (1989), and Yen et al. (1987) provide examples of the effects that customized tests nonrepresentative of the original tests can have on ability estimates. Tests customized by selecting specific content areas most relevant to the local curriculum and thus more familiar to the local students gave systematically higher ability estimates than estimates based on the full test; the same was not true when content was sampled representatively. The magnitude of the overestimates seemed dependent on the number of items deleted from the full test (Way et al., 1989).

Contradictory findings about the effects of differential instruction and textbooks were reported in a series of studies by Mehrens and Phillips. Neither the different textbook series used in grades 3 and 6 for reading and mathematics, nor the degree of instruction-test match based on teacher's ratings were found to impact standardized test scores significantly (Mehrens and Phillips, 1986). The small impact of these two curricular factors on unidimensionality was evaluated by a factor analytic method: the percentage of variance for the large first factor did not change noticeably, and the second factor remained relatively small across groups with different curricula (Phillips and Mehrens, 1987). In a third paper, they reported that item *p*-values and Rasch difficulty parameter estimates were similar across student groups using different textbook series (Mehrens and Phillips, 1987). The authors list a number of potential reasons for the lack of curricular impact, including the lack of power to detect differences, the precision of teacher's ratings of instruction-test match (Phillips and Mehrens, 1987), and emphasis on general competence versus specific details related to curricular objectives. Linn (1990) further comments on the findings by Mehrens and Phillips that their studies were done in elementary grades with widely used textbooks, in contrast to studies in higher grades that have qualitatively more different instructional experiences, and which found a demonstrated impact on test performance.

CONTEXT EFFECTS ON ITEM PARAMETER ESTIMATES

Apart from the content of items, the context in which they are presented also influences the estimates of item parameters. "A context effect occurs when a change in the test or item setting affects student performance" (National Research Council, 1999, p. 34). Masters (1988) considers (a) opportunity to answer, due to speeded tests, and fatigue, and (b) test wiseness as sources of differential item performance reflected in item parameters. Items that appear at the end of a test and items sensitive to test wiseness skills will favor students of higher ability and thus produce inflated discrimination parameters. Had the items appearing at the end of the test been presented earlier, they would have been attempted by more examinees and would have likely exhibited lower discriminations.

Item positioning effects were also examined by Kingston and Dorans (1984) for 10 item types on the GRE test. They found changes in the IRT equatings when those items were moved to a different position in the test. The effects were larger for the Quantitative subtest than for the Verbal, and even more profound for the Analytical. Practice and fatigue effects clearly depended on the location of an item and they seemed to interact with the type of the item: analytical, quantitative, or verbal. Similar findings are reported for an operational statewide testing program by Meyers et al. (2009). They found changes in Rasch difficulty values for items in a field test versus the final test depending on the change in the positioning of the items. In this study, the parameter estimates from the field testing are used for equating purposes and in a simulated investigation the authors conclude that if items are re-positioned in the live test form, as is done in practice, the equating "would benefit higher ability students and disadvantage lower ability students" (p. 57).

Yen (1980) reports that the location of an item in a booklet frequently affected the value of its difficulty parameter. Items placed at the end of a test had higher parameter estimates, i.e., were more difficult, than when presented at the beginning. Item location only partially explained parameter change in her paper. Similarity of item arrangements seemed to be another factor. The booklets "with the most similar item sequences tended to have more strongly related item parameter estimates than the booklets with the least similar item sequences" (p. 308). In contrast, Sykes and Fitzpatrick (1992) reported that changes of item location parameters were unrelated to changes in the booklet or the test position, and item type (tryout or scorable.)

A requirement for sound equating is that the equating function must be population invariant; the choice of (sub) populations to estimate the equating relationship between two tests should not produce large discrepancies (Dorans and Holland, 2000). Performance on items (common items when the NEAT design is used) drives equating functions. Differential item performance between groups would cause dependency of the equating relationship on the population. In a study of traditional equating methods, Kingston et al. (1985) looked at the equating functions between subgroups that took two different forms of the GMAT. Equated scores derived from the male and female subgroups were very similar, as were those derived from age or random subgroups. In a comparable study on the GRE and sex, race, field of study, level of performance and random subgroups, Angoff and Cowell (1985) also found support for the population invariance of equating.

A study by Cook et al. (1988) looked at a different kind of "context" and reached different conclusions. What was special about this study was that the samples employed to generate the equating transformation did not come from the same test administration, but from fall versus spring administrations of the test, and thus subgroups used to link the tests were dissimilar. Cook et al. (1988) demonstrated that when curriculum-related biology achievement tests were given to groups of students at different points in time after learning the content, item parameter estimates were unstable. For instance the correlation coefficient between the delta values[2] of 58 common items in two consecutive fall administrations was 0.99 as opposed to 0.79 between the fall and the spring administration values. Groups taking the test at different points in their coursework could not be considered as samples from the same population; recency of instruction, the time lapsed from when the material was taught, appeared to influence item parameter estimates. In addition, both linear and non-linear, including IRT, equating methods were not robust in such cases, giving very disparate scaled score summary statistics. In contrast, the statistics from equating forms administered at the same time period in consecutive years, i.e., fall of the first year and fall of the second year, were similar under all equating methods.

Disclosure of, or familiarity with items is another potential cause for changes in item location parameters. A security breach could have unpredictable effects on equated scores depending on whether the items exposed are common or not, and on the magnitude of the breach (Brennan and Kolen, 1987). A study simulating increasing levels of anchor item disclosure by randomly selecting and changing incorrect to correct responses resulted in an increasing drift in difficulty parameters as disclosure moved from low to moderate levels (Mitzel et al., 1999). More importantly, even at modest exposure levels, IRT equated score distributions altered considerably. Under a traditional Tucker linear equating method, Gilmer (1989) found modest effects on the passing rates on a certification test due to simulated item disclosure.

DESIRABLE CHARACTERISTICS OF COMMON ITEMS: SOME GUIDELINES

Common items provide the statistical means for equating test forms and making scores from different administrations of the same testing program comparable. Since tests need to be built to the same content and statistical specifications for the comparisons to be meaningful, the anchor items should proportionally reflect the specifications of the total test if they are to reflect group differences adequately (Brennan and Kolen, 1987; Cook and Petersen, 1987; Kolen and Brennan, 2004). For a non-random, common-item equating design Budescu (1985) noted that a high correlation between the anchor subtest and the two total tests is a necessary condition for stable and precise equating. Klein and Jarjoura (1985) argued that "it is important that the common items directly reflect the content representation of the full test forms. A failure to equate on the basis of content representative anchors may lead to substantial equating error" (p. 205).

Adequate numbers of common items need to be included to reduce random equating error, particularly in educational achievement tests, which are not strictly unidimensional. As a rule of thumb, at least 20 items or 20% of the length of a moderately long test should be used as anchors (Angoff, 1971; Kolen and Brennan, 2004). The longer the anchor test, the more reliable the equating will be (Budescu, 1985).

Additional precautions to avoid systematic influences on anchor items relate to their positioning, which should be approximately the same in the alternate forms (Cook and Petersen, 1987), and their presentation, which should be identical, i.e., without changing the text (Cassels and Johnstone, 1984) or the order of multiple-choice options (Cizek, 1994). The researchers making these caveats have shown that performance on items is sensitive to such variations. Noteworthy cases of context effects that appeared in large-scale testing programs come from the National Assessment of Educational Progress (NAEP) and the Armed Services Vocational Aptitude Battery (ASVAB). The large drop in proficiency observed between the 1984 and 1986 NAEP reading scores was termed the "NAEP reading anomaly" and was in part attributed to the dissimilar structure of the test booklets, the change in position of the common items and the different time available to respond to the common items in the two administrations (Zwick, 1991). In the late 1970s new forms of the ASVAB were introduced and the scaling was carried out under a single group design (no common items are used in this case). Examinees took both the old and the new form and they were able to distinguish between the two forms through the printing format and content of the forms; aware that only the old form scores would have been used for selection purposes, they were more motivated when responding to the old than the new forms. As a result, the passing scores estimated for the new form were much lower and individuals with lower skills were selected to enter the military (Maier, 1993; Kolen and Brennan, 2004). The latter example illustrates the importance of motivation in test taking and the implications when examinees can recognize if an item is common or not, especially when used in an external anchor set under a NEAT design, or in other designs such as when the equating is conducted with anchor items that have been pretested, or in a single group design.

Before they are judged appropriate for the equating process, anchor items must pass additional analyses after the administration of the tests to scrutinize their behavior, as reflected in item parameter estimates. Items that behave consistently over multiple administrations are appropriate for use in the test equating process. Items indicating anomalous parameter changes over time are likely to be rejected from the common item pool and treated as regular, non-common items. Hence, the number of anchors in the test should be sufficiently large to effectively complete the equating task after the rejection of some items.

Item response theory or classical item statistics may be used to examine whether embedded common items are functioning differentially for groups taking different test forms (Kolen and Brennan, 2004). A common approach in an IRT context is to scale two to-be-equated forms separately. Each calibration yields item parameter estimates that are used to generate a transformation to place the tests on a common scale. For example, a scatter plot of the common item's difficulty parameters estimated in the calibration

[2]Delta values are transformed proportion correct values defined as the inverse normal transformations of the *p*-values rescaled by multiplying by −4 and adding 13.

of the first form versus that of the second would show an approximately straight line under a satisfactory IRT model fit. Some random variation is expected, but clear outliers would suggest that the assumptions of the model are not met. Alternatively, the delta-plot method is used in practical situations to examine the volatility of equating item's difficulty values (Angoff, 1972); it is a simple and comprehensible graphical method for studying the item-by-group interaction, which makes use of the classical test theory difficulty indices, the *p*-values. The delta-plot flags outliers in a scatter plot of delta values of the common items obtained from two groups of examinees. The points that lie at a distance from the "cloud" of the majority of the points represent the common items whose *p*-values differ by an unexpectedly small or large amount. Those items are candidates for exclusion from the common-item pool. The delta-plot method is widely implemented because it is practical and does not involve IRT calibrations, which would be the case if IRT parameters were compared, and because it provides prima-facie evidence regarding anomalous changes in item difficulties across administrations.

ABERRANT BEHAVIOR OF COMMON ITEMS

Studies on the effect of simulated item parameter drift, i.e., the differential change in item parameters over time, on estimates of examinee proficiency have shown that individual ability estimates are fairly robust to non-common item parameter drift. When drift contaminates the common-item pool ability estimates are influenced. Stahl et al. (2002) simulated increasing levels of item parameter drift and observed the impact on Rasch estimates of examinee measures. Under conditions of simulated increase of item difficulties, and by varying the number of drifted items and the direction of the drift, ability estimates were robust; by setting a pass/fail cut-score, the majority of the misclassifications were within the 95% confidence band of the cut-score, "indicating that the misclassifications may be due purely to error of measurement and not to the effect of drift" (p. 8). Wells et al. (2002) applied a 2-Parameter Logistic IRT model, and in addition they examined what the effect was on ability estimates when drifted items were excluded from equating; they found little effect. Fitting a 3-Parameter Logistic model, Huang and Shyu (2003) simulated conditions of drift in the discrimination and difficulty parameters, varied the sample sizes and the percentage of the common items with drifted parameters and performed equating with or without the drifted items. When drifted items were excluded from equating, the equated scores did not differ much from the baseline scores; they did affect mean and passing scores when they were kept in the common item pool. Large increases in the difficulty parameters and when the drifted items constituted half of the common item pool had the more profound consequences, especially with a small sample size of 500.

The above studies on the impact of item drift on estimates of ability whether equated or not were all simulation studies, usually modeling unrealistically large drift in item parameters, typically in one direction and for a large number of items (common or not.) In reality, only a few common items demonstrate large changes in their item parameters. And those changes are not unidirectional; an item may become easier while another becomes more difficult, thus partially negating some of the effect of item drift. Therefore, in real situations the effects of change in estimated parameters will probably be much less. It remains true though that the effect, if any, on aggregate scores will still be much more profound than on individual examinee scores.

Using real data from four statewide assessment programs, Michaelides (2010) identified misbehaving common items across consecutive administrations via the delta-plot method. One to three items were flagged as aberrant, and the decision to include or discard them from the anchor pool had in two of the assessments non-negligible impact on aggregate statistics, such as overall score gains from one administration to the next and the percent of students above a cut score. In a simulation study, Sukin and Keller (2008) manipulated the inclusion or exclusion of one aberrant common item and concluded that there was no impact on the accuracy of examinee classification, but the percent of students over- and under-classified was different. Hu et al. (2008) found that including outlier common items with inconsistent difficulty parameters resulted in larger systematic error in the equated scores.

DISCUSSION AND CONCLUSIONS

Equating is an essential part of linking test scores across administrations and maintaining a common longitudinal scale for tracking trends in group performance over time, indicating how examinee cohorts perform compared to their counterparts. Within an accountability system that seeks to accurately capture student performance and growth, the treatment of misbehaving common items – whether arising from content or context effects – may introduce additional sources of error. Evidently, test development and administration procedures should follow suggested guidelines, such as those outlined in Section "Desirable Characteristics of Common Items: Some Guidelines" above, on how to develop and include common items in a testing program. As shown in this review, there is ample evidence from simulation studies, and investigations with real test data and practices that even minor departures from these guidelines could impact examinee performance on items.

The examination of the common items to determine their appropriateness for use in equating gives rise to a number of concerns regarding the valid interpretation and use of the equated test scores. Referring specifically to educational achievement tests, stable performance of students on tests is not necessarily a desirable property. Educational systems expect and seek progress in their student's learning, not only as a result of educational practice, but also through the implementation of innovative programs, reallocation of resources, new policies and reforms in the curriculum or administrative procedures. If an accountability system successfully encourages the reallocation of instructional resources, then some common items answered by different administration groups could appear anomalous possibly because they are indicating real effects: that the reform initiative has indeed made a difference and performance on the relevant items has changed. Consequently, their parameter estimates will not be invariant. When the items reflecting the results of the reallocation are removed on statistical grounds because of presumed violations of model assumptions the effects of the reform may be adjusted away.

The treatment of misbehaving common items pertains to the validity of the test. "If items that are found to be most sensitive to instruction are eliminated so that the IRT assumptions are better satisfied, there is a real danger that IRT will do more to decrease than to increase the validity of achievement test scores"

(Linn, 1990, p. 136). If items examining certain curricular domains – and particularly those domains at which recently implemented policies would be targeted – are deleted from the anchor pool, the content domain that the test is constructed to measure is redefined in ways that cannot be determined and limited to those items that do not disturb the model's assumptions. The removal of anomalously behaving common items is thus in discordance with the value placed by the system on that particular domain. As with the process of construction of good tests, items cannot be chosen merely on the basis of their psychometric attributes. Content is a legitimate consideration in deciding which items remain in the anchor.

It is not easy to provide strict guidelines on how to deal with common items flagged for differential behavior across two administrations. The content tested by a common item and its relevance to both the curriculum framework and actual instruction comes into the decision as to how to treat it, if it behaves in unexpected ways. As in the case of differential item functioning studies where an instance of an item functioning differentially for two groups does not necessarily imply that the item is biased and should be discarded from a test (Linn, 1993), finding a common item that fails to function consistently across administrations does not imply that it is inappropriate for equating. Content experts and test developers may offer plausible explanations for the differential behavior. If a context effect has, for example, been discovered, then it is probably legitimate to say that it is unrelated to the construct that the test is measuring; in this case, Miller and Fitzpatrick (2009) suggest that keeping it in the estimation of equating constants will result in equating error. However, as regards equating, even in obvious cases of discrepant performance due to irrelevant circumstances, discarding a common item is not as straightforward. Common items are chosen to meet certain content and statistical specifications, and to proportionally represent the properties of the total test. Discarding a common item might violate those guidelines and introduce a different kind of bias in the equating transformation.

Even though a judgmental decision about misbehaving common items is involved in the equating process, the practice can be improved in different ways. For example, the impact of including or excluding an item on the specifications and the content representation of the common-item pool can be examined. Knowing a common item's leverage, the decision on how to deal with it can be more informed. Studies that simulate realistic situations would provide insight on the importance of item characteristics that affect the leverage of outliers. Using such information together with the plausible causes of differential item behavior – content, context, or unidentifiable – the decision to keep or discard a common item can be more defensible. Future research may also look into alternative methodologies for flagging outlying common items in a more formal way than the delta-plot – see Michaelides (2008) for an example of treating this phenomenon as a case of differential item functioning.

Beyond the extent of influence that misbehaving common items can have on equating results, the reasons behind the unexpected behavior are worthy of investigation. If in fact there are very few, if any, outlying common items in achievement tests, then presumably those outliers are not caused by intentional curriculum and policy changes, but more likely by random, context, and, from an educational perspective, uninteresting events, such as changes in positioning, or accidental disclosure of items. A large-scale modification of the content standards for example would probably affect the behavior of many items over a long time period, while an extraordinary incident between two administrations that sensitized one of the examinee populations, or a change in the presentation of an item from one form to the next, would likely affect the behavior of the relevant item only. Nevertheless, outlying common items need not be the only studied items in this context. If a new policy is implemented between two administrations, then its effects can be examined on all common items, whether they exhibit large, small, or no differential behavior. Since educational policies often require longer time periods to implement and produce results, such future investigations, similar to Bock et al.'s (1988) study, could be longitudinal rather than just between two administrations, and would provide support for the link between educational policies and educational outcomes, reflected as changes in item (or item-cluster) performance.

ACKNOWLEDGMENTS

This research was partially supported by CRESST/UCLA, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education. Their support is gratefully acknowledged. Ideas and opinions stated do not necessarily represent CRESST/UCLA positions or policies. The author thanks Edward Haertel for comments on earlier drafts of the manuscript.

REFERENCES

Allen, N. L., Ansley, T. N., and Forsyth, R. A. (1987). The effect of deleting content-related items on IRT ability estimates. *Educ. Psychol. Meas.* 47, 1141–1152.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: AERA.

Angoff, W. H. (1971). "Scales, norms, and equivalent scores," in *Educational Measurement*, 2nd Edn, ed. R. L. Thorndike (Washington, DC: American Council on Education), 508–600. (Reprinted as Angoff, W. H. (1984). *Scales, Norms, and Equivalent Scores*. Princeton, NJ: Educational Testing Service).

Angoff, W. H. (1972). "A technique for the investigation of cultural differences," *Paper Presented at the Annual Meeting of the American Psychological Association*, Honolulu. (ERIC Document Reproduction Service No. ED 069686).

Angoff, W. H., and Cowell, W. R. (1985). *An Examination of the Assumption that the Equating of Parallel Forms is Population Independent* (RR-85-22). Princeton, NJ: Educational Testing Service.

Bock, R. D., Muraki, E., and Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *J. Educ. Meas.* 25, 275–285.

Brennan, R. L., and Kolen, M. J. (1987). Some practical issues in equating. *Appl. Psychol. Meas.* 11, 279–290.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *J. Educ. Meas.* 22, 13–20.

Cassels, J. R. T., and Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *J. Chem. Educ.* 61, 613–615.

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educ. Psychol. Meas.* 54, 8–20.

Cook, L. L., Eignor, D. R., and Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *J. Educ. Meas.* 25, 31–45.

Cook, L. L., and Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal

circumstances. *Appl. Psychol. Meas.* 11, 225–244.

Dorans, N. J., and Holland, P. W. (2000). Population invariance and the equitability of tests: basic theory and the linear case. *J. Educ. Meas.* 37, 281–306.

Gilmer, J. S. (1989). The effects of test disclosure on equated scores and pass rates. *Appl. Psychol. Meas.* 13, 245–255.

Hambleton, R. K., Swaminathan, H., and Rogers H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications, Inc.

Hanson, B. A., and Feinstein, Z. S. (1997). *Application of a Polynomial Loglinear Model to Assessing Differential Item Functioning for Common Items in the Common-Item Equating Design (ACT Research Report Series No. 97-1).* Iowa City, IA: ACT Inc.

Hattie, J., Jaeger, R. M., and Bond, L. (1999). Persistent methodological questions in educational testing. *Rev. Res. Educ.* 23, 393–446.

Hu, H., Rogers, W. T., and Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Appl. Psychol. Meas.* 32, 311–333.

Huang, C. Y., and Shyu, C. Y. (2003). "The impact of item parameter drift on equating," *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*, Chicago.

Klein, L. W., and Jarjoura, D. (1985). The importance of content representation for common-item equating with non-random groups, *J. Educ. Meas.* 22, 197–206.

Kingston, N., Leary, L., and Wightman, L. (1985). *An Exploratory Study of the Applicability of Item Response Theory Methods to the Graduate Management Admissions Test* (RR-85-34). Princeton, NJ: Educational Testing Service.

Kingston, N. M., and Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Appl. Psychol. Meas.* 8, 147–154.

Kolen, M. J., and Brennan, R. L. (2004). *Test Equating: Methods and Practices,* 2nd Edn. New York: Springer.

Linn, R. L. (1990). Has item response theory increased the validity of achievement test scores? *Appl. Meas. Educ.* 3, 115–141.

Linn, R. L. (1993). "The use of differential item functioning statistics: a discussion of current practice and future implications," in *Differential Item Functioning*, eds P. W. Holland and H. Wainer (Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers), 349–364.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Maier, M. H. (1993). *Military Aptitude Testing: The Past Fifty Years* (DMDC Technical Report 93–007). Monterey, CA: Defense Manpower Data Center.

Masters, G. N. (1988). Item discrimination: when more is worse. *J. Educ. Meas.* 25, 15–29.

Michaelides, M. P. (2008). An illustration of a Mantel–Haenszel procedure to flag misbehaving common items in test equating. *Pract. Assess. Res. Eval.* 13. Available online: http://pareonline.net/getvn.asp?v=13&n=7.

Michaelides, M. P. (2010). Sensitivity of equated aggregate scores to the treatment of misbehaving common items. *Appl. Psychol. Meas.* 34, 365–369.

Mehrens, W. A., and Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *J. Educ. Meas.* 23, 185–196.

Mehrens, W. A., and Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *J. Educ. Meas.* 24, 357–370.

Meyers, J. L., Miller, G. E., and Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Appl. Meas. Educ.* 22, 38–60.

Miller, A. D., and Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *J. Educ. Meas.* 25, 205–219.

Miller, G. E., and Fitzpatrick, S. J. (2009). Expected equating error resulting from incorrect handling of item parameter drift among the common items. *Educ. Psychol. Meas.* 69, 357–368.

Mitzel, H. C., Weber, M. M., and Sykes, R. C. (1999). "Test item disclosure: how much difference does it really make?," *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*, Montreal, Canada.

National Research Council. (1999). "Embedding questions: the pursuit of a common measure in uncommon tests," in *Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education*, eds Committee on Embedding Common Test Items in State and District Assessments, D. M. Koretz, M. W. Bertenthal, and B. F. Green (Washington, DC: National Academy Press).

Phillips, S. E., and Mehrens, W. A. (1987). Curricular differences and unidimensionality of achievement test data: an exploratory analysis. *J. Educ. Meas.* 24, 1–16.

Rupp, A. A., and Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educ. Psychol. Meas.* 66, 63–84.

Skaggs, G., and Lissitz, R. W. (1986). IRT test equating: relevant issues and a review of recent research. *Rev. Educ. Res.* 56, 495–529.

Stahl, J., Bergstrom, B., and Shneyderman, O. (2002). "Impact of item drift on person measurement," *Paper Presented at the Annual Meeting of the American Educational Research Association*, New Orleans, LA.

Sukin, T., and Keller, L. (2008). "The effect of deleting anchor on the classification of examinees," in *NERA Conference Proceedings 2008*. Downloaded on February 20, 2010 from http://digitalcommons.uconn.edu/nera_2008/19

Sykes, R. C., and Fitzpatrick, A. R. (1992). The stability of IRT b values. *J. Educ. Meas.* 29, 201–211.

von Davier, A. A., Holland, P. W., and Thayer, D. T. (2004). *The Kernel Method of Test Equating.* New York, NY: Springer.

Wainer, H. (1999). Comparing the incomparable: an essay on the importance of big assumptions and scant evidence. *Educ. Meas. Issues Pract.* 18, 10–16.

Way, W. D., Forsyth, R. A., and Ansley, T. N. (1989). IRT ability estimates from customized achievement tests without representative content sampling. *Appl. Meas. Educ.* 2, 15–35.

Wells, C. S., Subkoviak, M. J., and Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Appl. Psychol. Meas.* 26, 77–87.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *J. Educ. Meas.* 17, 297–311.

Yen, W. M., Green, D. R., and Burket, G. R. (1987). Valid information from customized achievement tests. *Educ. Meas. Issues Pract.* 6, 7–13.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educ. Meas. Issues Pract.* 10, 10–16.

Conflict of Interest Statement: The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 February 2010; paper pending published: 26 February 2010; accepted: 22 September 2010; published online: 15 October 2010.

Citation: Michaelides MP (2010) A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating. Front. Psychology 1:167. doi: 10.3389/fpsyg.2010.00167

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.